DOCUMENT RESUME

ED 335 421 TM 017 141

AUTHOR	Resnick, Lauren B.; Resnick, Daniel P.
TITLE	Tests as Standards of Achievement in Schools.
INSTITUTION	Center for Technology in Education, New York, NY.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Oct 89
NOTE	22p.; Paper presented at the Invitational Conference of the Educational Testing Service (New York, NY, October 28, 1989).
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Standards; *Achievement Tests; Educational Assessment; Educational Change; *Educational Improvement; Elementary Secondary Education; Standardized Tests; *Student Evaluation; *Testing Problems; Test Use; *Thinking Skills
IDENTIFIERS	Curriculum Based Assessment; Performance Based Evaluation

ABSTRACT

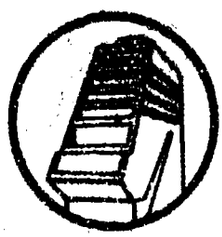
The question of whether tests can be both curriculum-neutral and effective means of monitoring and motivating educational practice is discussed. Educational reform and testing are intimately linked, as tests are widely viewed as a means of educational improvement. Tests/assessments influence educator behavior by stimulating them to assure that their students perform well. Tests/assessments used for public accountability or program evaluation purposes affect the curriculum. A new vision of education--a thinking-oriented curriculum (TC) for all students--is considered, in which education focuses on higher-order abilities, problem solving and thinking, and the ability to go beyond the routine and exercise personal judgment. Current tests that are inimical to a TC are discussed. To assess the extent to which decomposition and decontextualization--two key assumptions underlying standardized testing--permeate today's achievement tests, reading comprehension, language, and mathematics test batteries that are widely used in educational assessment are analyzed. Standardized tests fare badly when judged against the criterion of assessing and promoting a TC. They embody a view of education that defines knowledge and skill as a collection of bits of information and they demand fast non-reflective replies. Alternative performance assessments for a TC, including open-ended writing examinations (essays) and portfolio assessments, help release educators from the pressure toward fractionated low-level forms of learning that are rewarded by most current tests, and they also set positive standards for an educational system that strives to cultivate thinking. Tied to curriculum and designed to be taught to, performance assessments can be essential tools for raising authentic educational achievement. A 25-item list of references is included. (RLC)

Tests as Standards of Achievement in Schools

Lauren B. Resnick
Learning Research and Development Center
University of Pittsburgh

Daniel P. Resnick
Carnegie Mellon University

LEARNING RESEARCH AND DEVELOPMENT CENTER



University of Pittsburgh

Tests as Standards of Achievement in Schools

Lauren B. Resnick
Learning Research and Development Center
University of Pittsburgh

Daniel P. Resnick
Carnegie Mellon University

October 1989

Essay prepared for the Educational Testing Service Conference
The Uses of Standardized Tests in American Education
New York

This is an adaptation of an extended paper by L. B. Resnick and D. P. Resnick, Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Future assessments: Changing views of aptitude, achievement, and instruction. Boston: Kluwer Academic Publishers.

The Uses of Standardized Tests in American Education

Proceedings of the 1989 ETS Invitational Conference

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

Tests as Standards of Achievement in Schools

LAUREN B. RESNICK
University of Pittsburgh
and
DANIEL P. RESNICK
Carnegie Mellon

In America, educational reform and testing are intimately linked. Test scores signal the need for reform, as evidenced by the attention paid to declining scores on college entrance exams and standardized tests, to Americans' weak ranking in international comparisons, and to the percentages of students performing poorly on certain kinds of items in our national assessments.

Tests are also widely viewed as instruments for educational improvement. Calls for better performance by American schools are almost always accompanied by increases in the amounts of testing done in the schools. New tests — and more active scrutiny of tests already in place — are frequently prescribed, both as a source of information for a concerned public and as a form of quality control and an incentive for better performance by educators and students.

At the same time, the rhetoric surrounding the introduction and interpretation of assessment programs often suggests that tests are not meant to influence curricula and teaching directly. This rhetoric of curriculum-neutral tests accords well with American traditions of local control over education, producing a profound and continuing resistance to any attempt to impose a curriculum from outside a school district. If tests and assessments are considered curriculum-neutral — not geared to any particular instructional program and not imposing any particular set of goals or practices — they can be incorporated easily into the ideology of local educational control. If tests are recognized as guiding

---

This is an adaptation of an extended paper by L. B. Resnick and D. P. Resnick, Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (eds.), *Future assessments: Changing views of aptitude, achievement, and instruction*. Boston, MA: Kluwer Academic Publishers.

or constraining the curriculum, they become problematic within our educational ideology.

How are educators and the public at large to make sense of this discussion? Can tests be both curriculum-neutral and an effective means of monitoring and motivating educational practice? Are tests only *indicators* (see Fuhrman, 1988; Murnane & Raizen, 1988) of how well schools are performing, without a direct influence on teaching? Or do they influence the school curriculum? Answers to these questions lie not in a theory of how tests *should* be used, but in a dispassionate analysis of the ways in which assessments function as elements in the social system of schooling.

In assessing complex systems, we often aim for indirect indicators of desired properties rather than for direct measurements. For example, to obtain an indicator of the amount of ambient heat in the air and thus how comfortable a room is for its occupants, we examine the height of mercury in a confined column and take a numerical temperature reading. We do not really care about the height of the mercury, however. What we normally care about is the physical comfort of people in the room. If we were to measure comfort directly, we would examine to what extent people were sweating, shivering, or showing other signs of physical discomfort, or ask them to rate their degree of comfort. One reason for using the temperature indicator is its unobtrusive character; taking a thermometer reading does not change the degree of comfort in the room.

Discussions of educational testing and educational standard setting often use the language of indicators. Educational tests, however, do not share the unobtrusiveness of indicators used in other measurement systems. We cannot place a "test thermometer" in a classroom without expecting to change conditions in the room significantly. Because educational tests are used in a social rather than a physical system, measurements that are made known to actors in the system can be expected to affect future actions. Molecules of air are not prompted to produce a particular temperature, but teachers and school principals can be motivated to produce test scores in an acceptable range. Any educational assessment that receives publicity will stimulate educators to assure that their students perform well on that assessment.

This power of tests and assessments to influence educator behavior is precisely what makes them potent tools for improving educational standards. Tests are introduced not just to provide neutral indicators of the education system's performance, but also in the hope of upgrading

curriculum, teaching, and academic performance. There is considerable evidence that this strategy works, insofar that it produces a rise in test scores. Even in school districts with an official policy against teaching to tests, considerable attention in the press or elsewhere to test scores causes teachers to adapt their teaching to the tests. Often, the effects of this adaptation become visible only after a new test, with different emphases, is adopted by or imposed on the district. Test scores in grade equivalent or other comparative terms then drop.

To account for this recurrent observation, several analysts have focused attention on the extent to which test items and curriculum activities correspond. When overlap between test and curriculum is high, test scores are high; when overlap decreases, so do test scores (Leinhardt & Seewald, 1981). School districts and teachers try to maximize overlap by choosing tests that match their curriculum. When they cannot control the tests, they strive for overlap by trying to match curriculum to the tests; i.e., by "curriculum alignment." In the first few years of a test's use, overlap increases as the curriculum is aligned. When a new test is imposed, overlap suddenly decreases, because the curriculum cannot change as quickly as the test can.

Some educators (e.g., Popham, 1987) have argued that the process of curriculum alignment is a favorable one and should be publicly encouraged and supported with tools to make teaching to the tests easier and more reliable. Evidence from parts of the country in which measurement-driven instruction has been adopted indicates that such instruction can, by focusing attention on a small set of desired objectives, improve performance on a particular set of test items. But this apparent success may actually mask stagnation or even decline in the kind of school achievement that is the real goal of educational reform today. For when the stakes are high — when school ratings and budgets or teacher salaries depend on test scores — efforts to improve performance on a particular assessment instrument seem to drive out most other educational concerns.

Shepard (1988) has studied this process in Texas, where teachers were given materials suggesting instructional strategies for each of the objectives on the state assessment. The strategies, which were specific to the test item forms, promised teachers who used them high test performance for their students, because the curricula would be perfectly aligned with the tests. Commercially sold programs to help students learn test-taking skills, Shepard found, are also closely tied to specific item types that appear on the major standardized test batteries. Under these condi-

tions, the range of skills taught is restricted, and slight variants in format that might be equally valid ways of exercising a skill are ignored, in favor of drilling students on the precise item types they will encounter on the tests. Other investigators (e.g., Cohen, 1987; Kellaghan, Madaus, & Airasian, 1980; Romberg, Zarinnia, & Williams, 1989) have further documented the tendency of high-stakes tests to progressively restrict curricular attention to the objectives that are tested and to the particular item types that will appear on the tests.

Whether we like it or not, what is taught and what is tested are intimately related. Public accountability systems will eventually influence what is taught and how it is taught in the schools. We must think of every test or assessment used for public accountability or program evaluation purposes as an instrument that will affect the curriculum. For those who would use tests as a means of monitoring school achievement, three principles may serve as guidelines.

- **You get what you assess.** Educators will teach to tests if the tests matter in their own or their students' lives, making tests potential tools in educational reform. Tests must be carefully crafted to sample directly those educational performances that are valued. Indicators of desired goals, no matter how well they may correlate with the truly desired outcome, are not good public accountability measures. For example, multiple-choice tests can be designed to correlate very highly with written composition grades. Such tests are good indicators of composition skill, but if we put many of them into the testing system, we must expect children to practice answering multiple-choice questions. In contrast, if we put debates, discussions, essays, and problem solving into the testing system, children will spend time practicing those activities.

- **You do not get what you do not assess.** What does not appear on tests tends to disappear from classrooms in time. If the goals of solving complex problems or writing extended essays are educationally important, those activities need to be sampled directly in an assessment program aimed at encouraging improved instruction. It is not sufficient to test the basics (a common strategy in today's assessment programs) or to assume that preparing for the tests will take minimal time and that teachers can then go on to other higher-order abilities.

- **Build assessments toward which you want educators to teach.** This principle follows directly from the first two and lies at the heart of the matter. Assessments should be designed so that when teachers do the natural thing — that is, prepare their students to perform well — they exercise the kinds of abilities and develop the skills and knowledge that are the real goals of educational reform. This principle assumes that what is in the assessment will be practiced in the classroom in a similar form. It proposes the central question for any assessment exercise: "Is this what we want students to be doing with their educational time?"

The Challenge of the Thinking Curriculum

By placing curriculum at the heart of testing decisions, these principles assert that tests must be chosen to assess directly and, thereby, promote the goals considered most central and important in education. Judged in these terms, most current tests are severely wanting. They are tuned to a curriculum of the past, one not suited to today's social and economic conditions.

In the last several years, a new vision of education has emerged, fueled partly by the needs of a changing economy and partly by recent research on learning and cognition. According to this view, education must focus on higher-order abilities, on problem solving and thinking, on the ability to go beyond the routine and to exercise personal judgment. Analyses of how technology is affecting the workplace and communication point to the need for workers at all levels to understand the technical systems they use, so they can participate in dispersed management systems requiring judgment and decision making (Resnick, 1987b; Scribner, 1984; Zuboff, 1988). Furthermore, working conditions are likely to change several times during an individual's work life, requiring a capacity for adaptive learning. Employers are finding that students now leaving high school are not prepared to function well in the work environments they enter. Like colleges, employers are calling on schools to provide educational programs that enable graduates to reason and think, not just perform routine operations.

A thinking-oriented curriculum for all constitutes a significant new educational agenda. Although it is not new to include thinking, problem solving, and reasoning in *some* students' school curriculum, it is new to include it in *everyone's* curriculum. It is new to aspire seriously to make thinking and problem solving regular aspects of the school program for the entire population, even minorities, non-English speakers, and eco-

nomically disadvantaged children. Developing educational programs that assume all individuals, not just the elite, can become competent thinkers is a new challenge.

To meet this challenge, thinking must pervade the entire school curriculum for all students, from the earliest grades. One of the most important findings of recent research is that the kinds of mental processes associated with thinking are not restricted to an advanced or higher-order stage of mental development (Resnick, 1987a; Resnick & Klopfer, 1989). Instead, thinking and reasoning are intimately linked to successful learning of even elementary levels of reading, mathematics, and other school subjects.

The traditional view — that the basics can be taught as routine skills, with thinking and reasoning to follow later — can no longer guide our educational practice. We know that one cannot effectively memorize without organizing knowledge. Facts acquired without structure and rationale disappear quickly. Children cannot understand what they read without making inferences and using information that goes beyond the written text. They cannot become good writers without engaging in complex planning and self-evaluation. It is not possible for them to learn basic math skills well if they only memorize rules for manipulating written numerical symbols. Science learning requires students to build explanatory theories they can believe. All of this means that the skills we are accustomed to calling *higher-level* are part of the most basic competencies.

The thinking curriculum does not imply that instruction in processes of reasoning is a substitute for acquiring substantial knowledge. Instead, recent research teaches us to be highly respectful of knowledge as a requirement for good thinking. People who know more about a topic reason more profoundly about it than people who know little about it.

But the knowledge required for good thinking can only be acquired through processes of thinking. For concepts and organizing knowledge to be mastered, they must be used generatively — that is, they have to be called on over and over again, as ways to link, interpret, and explain new information. Education requires an intimate linking of thinking processes with knowledge content. This in turn calls for a reorganization of schooling, so that thinking suffuses the curriculum and is demanded in every subject.

Current Tests: Inimical to the Thinking Curriculum

In light of the demands of the thinking curriculum, most current tests work against the reforms required in our educational system. Testing practice remains essentially unchanged from the era in which it was considered enough for schools to teach mastery of routine skills — doing simple computations, reading predictable texts, reciting civic or religious codes. Goals such as interpreting unfamiliar texts, constructing convincing arguments, understanding complex systems, developing approaches to problems, or negotiating problem resolution in a group were reserved for an elite.

Two key assumptions, *decomposability* and *decontextualization*, underlie standardized testing technology and practice. These assumptions were compatible with the routinized skill goals of the past and with the psychological theories of the first part of this century. They are, however, incompatible with thinking goals for education and with what we know today about the nature of human cognition and learning.

Decomposability. Psychological theories of the 1920s assumed that thought could best be described as a collection of independent pieces of knowledge. That assumption can be clearly recognized in the work of psychologist Edward L. Thorndike, which profoundly influenced instruction and testing from the 1920s onward. In 1922, Thorndike published *The Psychology of Arithmetic*, in which he showed how the content of the elementary school arithmetic curriculum could be analyzed as a collection of "bonds" between stimuli and responses. Thorndike proposed that the task of arithmetic instruction was to exercise the bonds that comprise arithmetic, rewarding correct responses and "stamping out" incorrect ones. Under this model, students who acquired all of the bonds could be said to know arithmetic completely. Students who acquired fewer bonds, or who learned them to a less reliable criterion of performance, could be said to have measurably less arithmetic knowledge.

With this analysis of the nature of arithmetic knowledge and skill, constructing efficient, objective tests posed little problem. It was impractical to test all possible bonds, but samples could easily be tested on any given occasion. If neither students nor teachers knew exactly which arithmetic facts or procedures would appear on a given test, they had to practice all of them, or all in a given subsection of a curriculum, in order to perform well. According to Thorndike, performance on a collection of specific items constituted a valid indicator of how much of the whole

body of arithmetic a child knew.

This kind of sampling of independent bits of knowledge and skills, now enhanced by much more sophisticated psychometric tools and theories, remains the basic strategy for standardized testing. But the decomposability assumption has been seriously challenged by recent cognitive research, which recognizes that complicated skills and competencies owe their complexity not just to the number of components they engage, but also to interactions among the components and heuristics for calling upon them.

Complex competencies, therefore, cannot be defined just by listing all their components. Information-processing theories of cognition (e.g., Anderson, 1983; Newell & Simon, 1972) analyze cognitive performances into complexes of rules, but performances critically depend on interactions among those rules. Each rule can be thought of as a component of the total skill, but the rules are not defined independently of one another. The competence of a problem-solving system thus depends on how the complex of rules acts together. Other cognitive theories, which stress the role of structured knowledge and organizing principles in learning and thinking, move even further from the decomposition assumption.

All of this suggests that efforts to assess thinking and problem-solving abilities by identifying separate components of those abilities and testing them independently interferes with effectively teaching such abilities. Assessing separate components encourages exercises in which isolated components are practiced. But since the components do not add up to thinking and problem solving, students who practice only the components are unlikely to learn real problem solving or interpretive thinking.

Decontextualization. The second major assumption built into standardized tests asserts that each component of a complex skill is a fixed entity that will take the same form wherever it is used. If students know how to distinguish a fact from an opinion, for example, they can do so under all conditions of argument and debate, in all knowledge contexts. Under this assumption, it makes sense to select key critical thinking skills for decontextualized practice in school.

But the assumption no longer appears valid. Recent developments in the epistemology and philosophy of science (e.g., Lakatos, 1978; Toulmin, 1972) show that there is no absolute line between fact and theory, data and interpretation. Instead, what is counted as fact depends on tools and instruments with built-in theories, and on communally accepted methods for deciding among competing assertions. Thus, history and literature, as well as science and mathematics, must be understood as interpretive domains in which knowledge and skill cannot be detached from their contexts of practice and use. Educationally, this suggests that we cannot teach a skill component in one setting and expect it to be applied automatically in another. We cannot validly assess a competence in a context very different from that in which it is practiced and used. In writing, for example, decontextualized editing exercises, a common element in standardized test batteries, do not reveal what people do when they edit their own work. If we are trying to educate people who can craft phrases and sentences to convey intended meanings, editing tests set a false direction. Such decontextualization does violence to the kinds of abilities we seek.

To gauge the extent to which the decomposition and decontextualization assumptions permeate today's achievement tests, we examined the standardized test batteries widely used in educational assessment by individual school districts and in state assessments of educational quality as part of mandated testing programs.

Reading comprehension. Reading comprehension tests generally present short passages (an average of 250-350 words in the grades 8-11 tests we analyzed), together with multiple short questions. In asking for bits of information rather than interpretation of an extended passage, these tests reflect the decomposability assumption, treating knowledge and skill as accumulations of isolated pieces of information and not as coherent, interactive systems. Furthermore, the tests encourage quick finding of answers rather than reflective interpretation. The tests we examined allow students an average of five to six minutes to read a series of brief passages and answer five to eight questions about each. Although the tests require a degree of textual interpretation, their isolated questions rarely examine how students interrelate parts of the text and do not require justifications that support the interpretations. The nature of the questions and the speed with which they must be answered do not invite the kind of reflection and elaboration demanded by the thinking curriculum.

These tests tacitly convey a definition of reading as perusing short passages to answer other people's questions. Furthermore, the test format suggests that the answers to these questions are already known by the person asking them. Under these conditions, reading comprehension appears to be a matter of finding predetermined answers, not interpreting the written word. Children who practice reading mainly in

the form in which it appears on the tests have little exposure to the demands and reasoning possibilities of the thinking curriculum.

Language. The other standardized subtests devoted to language engage students in even less contextualized and extended thinking than the comprehension tests do. Vocabulary tests present decontextualized words in questions that must be answered at a rate of two or three per minute if the whole test is to be completed. Spelling tests usually contain items in which the student selects a proper spelling from among a set of misspellings — again at a rate of two or three per minute. There are various subtests on language usage, mechanics, expression, and punctuation. The items involve recognizing errors and choosing (not producing) corrections, usually at the rate of two or three items per minute. Students who practiced exercises similar to those that fill the standardized language tests would not learn to write coordinated, coherent prose. They might not even learn to write locally correct prose or to use a wide range of vocabulary, for there is good evidence that recognizing other people's errors and choosing the correct alternatives are not the same processes as those needed to produce good written language. These tests carry the decontextualization assumption to the extreme.

Mathematics. On the whole, the mathematics portions of the standardized tests fare even less well than the reading portions on the criteria laid out in this paper. All the tests contain major sections in which arithmetic computations are to be performed at the rate of one or two problems per minute. These are, perhaps, reasonable assessments of computational fluency; in any case, they do not claim to assess mathematical reasoning.

Much more disturbing are the subtests aimed at assessing mathematical concepts and problem solving. These, too, consist of many short, unrelated items, usually presenting problems to be solved at the rate of about one per minute. Recent publications of the National Council of Teachers of Mathematics (1989) and the Mathematical Sciences Education Board (National Research Council, 1989) establish standards for a conceptually oriented thinking curriculum in mathematics and call for extended mathematical reasoning, including problems that can be attacked by several different methods. None of the standardized mathematics tests that we examined even approximates these standards.

In summary, the standardized tests[1] fare badly when judged against the criterion of assessing and promoting a thinking curriculum. They embody a definition of knowledge and skill as a collection of bits of information, and they demand fast, nonreflective replies. The tests and the classroom practices that might be used to prepare for them suggest to students a view of knowledge counter to what the thinking curriculum seeks to cultivate: If you do not know an answer immediately, there is no way of arriving at a sensible response by thought and elaboration. Although some reading comprehension items demand interpretation of and inference from the text, questions are usually presented as isolated and unconnected with each other, with no hint that interpreting a text might involve an extended line of reasoning. The multiple-choice format, furthermore, reinforces the idea that someone else already knows the answer to the question, so original interpretations are not expected; the task is to find or guess the right answer, rather than to engage in interpretive activity.

Alternative Assessments for the Thinking Curriculum

Although the tests most widely used to assess achievement are unfriendly to the goals of the thinking curriculum, it is possible to develop assessments that will actually enhance thinking and reasoning abilities when teachers gear their instruction to the tests. Several states have recently added to their assessment batteries a writing examination, in which students produce essays that are graded by panels of judges to yield quantitative scores. A similar writing assessment is now included in the National Assessment of Educational Progress (NAEP).

These writing tasks begin to meet the criteria laid out in this essay for educationally appropriate assessment. If students engaged regularly in the type of activities found in the assessment, they would be practicing writing in an authentic form. It is possible to teach to these tests without destroying their educational validity. They represent potentially powerful tools of educational improvement, because their presence in the assessment system will actively encourage educators to provide significant amounts of writing practice in the curriculum.

[1] Although our detailed analysis was limited to widely used, commercially developed tests, most state-developed tests, as well as NAEP, use similar kinds of items and are subject to the same general critique.

The adoption of open-ended writing assessments by several states and by NAEP marks an important change in assessment policy. National and state testing agencies are now recognizing that open-ended responses can be scored with sufficient reliability to provide data to the public and to the educational system on the quality of learning. The use of writing assessments has shown the feasibility of using complex, integrated performances, rather than series of isolated questions, in public accountability systems. Their use has also shown that it is possible to derive reliable, credible quantitative measures from judgments of these products rather than from precoded correct answers. The successful use of writing assessments as part of public accountability testing opens the way for a much wider variety of new assessment methods, methods that are more compatible with the nation's aspiration to education for thinking.

The essay assessments are an example of a broader category of assessments, often called *performance assessments*. A performance assessment uses direct judgments and evaluations of performances, rather than indirect indicators of competence. Performance assessments are widely used in the arts and athletics. At the Olympics, for example, performances with no direct competition (such as diving and gymnastics) are rated by judges, and the pooled ratings are used to decide who wins medals. In music competitions, pianists or violinists perform a prescribed or self-selected repertoire; these performances are rated by judges, and again pooled ratings determine the winners. Ratings are often made on several separate dimensions, as well as on overall, global performance, and there may be complex formulas for weighting the different judgments.

A variant of the performance assessment is the *portfolio assessment* (see Gardner, in press). This method, frequently used in the visual and plastic arts and other design fields, requires individuals to collect their work over a period of time, select a sample of the collection that best represents their capabilities, and submit this portfolio to a jury or panel of judges.

Although best developed in the arts and athletics, performance and portfolio assessments are adaptable to other domains of knowledge and skill. The simplest form of performance assessment is the written essay, which can be used not only to assess writing skill, but also to assess knowledge of issues and ideas within a subject. Special forms of essay examinations also yield evidence of students' ability to carry out investigations and analyses of data. For example, the Advanced Placement Test in History contains a document-based question, in which students must analyze a set of documentary sources to answer an interpretative question. In England, where open-ended essay examinations have always been part of the graduation examinations taken by students at 16 and 18 years of age, there have been recent experiments with the use of extended project reports as part of the formal assessment system. The Manchester Joint Matriculation Board's Engineering Science Examination, for example, includes both experimental investigations and extended applied projects as part of the assessment portfolio (Joint Matriculation Board, 1982). Candidates conduct experiments or investigative projects on topics such as measuring strain in a model suspension bridge, estimating the volume of water flowing in a river, and designing and building a device for evaluating sound insulating properties of common building materials over the course of several months. They then submit reports on their plans, execution, outcome, and interpretation to the examining board. These reports are rated on each of several criteria (e.g., theoretical understanding, planning, design, use of procedures and equipment, possible alternative solutions considered, quality of the written report), and ratings are averaged to yield an overall grade for each candidate.

The Advanced Placement Tests and the Engineering Science Examination just described are equivalent to first-year college course examinations and are intended for only a fraction of the secondary school population. Performance assessments are, however, equally suitable for younger and broader populations of students. In this country, the National Assessment of Educational Progress has studied the feasibility of using open-ended exercises to assess higher-order thinking in science and mathematics at grades 3, 7, and 11 (Blumberg, Epstein, MacDonald, & Mullis, 1986; National Assessment of Educational Progress, 1987). These assessment exercises included written responses to problems; "station activities," in which individual students used equipment to investigate a phenomenon and then answered open-ended questions about it; and complete experiments that students designed, carried out, and reported orally. Some of the exercises were graded on the basis of students' written answers — their *products*. Others required observers to rate the *processes* students revealed as they worked. In both cases, graders had to be trained to apply common criteria and standards.

Videotaping of performances, a technology now inexpensive and reliable enough for widespread use, could in the future substantially simplify grading when direct observation is necessary. Indeed, the ease

of videotaping makes possible a wide variety of assessments in which one examiner interviews a student in a manner designed to probe understanding and thinking abilities, and a different set of graders scores the student's performance. We are experimenting with this form of assessment in a primary grades mathematics project (Resnick, Bill, & Lesgold, 1989). In one kind of assessment interview, a child is asked to solve an arithmetic problem and to explain his or her solution. In some of our interviews, the child is then shown an alternative solution and asked whether it too could be correct, and, if so, how two different solutions can yield the same answer. Performances of this kind can be graded on multiple criteria, such as the sophistication of the procedure, the completeness of the explanation, whether the child explains the solution conceptually or only procedurally, and whether the child produces the explanation spontaneously or needs to be questioned or prompted by the examiner. These multiple ratings could easily be reduced to reliable single scores in order to use these interview results for public accountability purposes.

If widely adopted as part of the public accountability assessment system in education, performance assessments (including portfolio methods) could not only remove current pressures for teaching isolated collections of facts and skills, but could also provide a positive stimulus for introducing more extended thinking and reasoning activities in the curriculum. The adoption of performance assessment methods would require educators to describe the kinds of thinking performances desired and the criteria of excellent performance much more precisely.

Once introduced, performance assessments would also assure a continuing forum for refining objectives and criteria for the thinking curriculum. There is good evidence for this in the experience of states that have used writing assessments for a few years. In those states, educators are beginning to discuss whether the kinds of essays students are asked to write reflect adequately the educational goals for writing. What is most striking is that the debates are primarily about curriculum and learning goals, not about techniques of assessment.

Using performance assessments as part of public accountability programs would require panels of trained judges to evaluate students' performance using specific criteria, ensuring sufficient agreement among the judges and reliable, unbiased scores from a set of individuals. Strategies for such training have been developed by various groups experienced in open-ended performance assessments in education. Teachers who have served on judging panels often report that the

training and review sessions help them develop and refine criteria for their own classroom work. Recognizing this, some school districts involved in performance assessment programs are discussing possibilities for using the training sessions as part of their staff development programs. Although it is too early for definitive evidence, experience to date suggests that teacher participation in judging and grading performance assessments can serve an important role in the general upgrading of educational standards.

One frequently raised objection to performance assessments is their high cost relative to the machine scorable tests now used. Performance assessments are more costly than current precoded tests, because multiple judges are needed every time an assessment is given. In public accountability assessment, in which an educational *system*, not individual students, is being evaluated, the costs of a full assessment *program* can be kept within tolerable bounds by testing less frequently and sampling more lightly than is currently done in many mandated testing programs.

Various schemes for light sampling have been developed, including methods that examine only some students, and those that examine all students in a given grade in which individuals take only a portion of the examination. Determining what justifiable inferences about student competencies can be made from different sampling procedures requires considerable technical, statistical sophistication. These issues are receiving continuing attention by certain states, by NAEP, and by commissions and study panels devoted to questions of assessment practice.

Thus, a scientifically sound basis exists for controlling costs by reducing the amount of testing, rather than by insisting on cheap-to-administer, precoded forms. To benefit from light sampling methods, states and other educational authorities will have to resist the temptation to combine accountability assessment with other testing functions requiring data on individual students, such as instructional diagnosis or student selection and certification. Accountability assessments should not attempt to offer diagnostic or other instructional management information. Such efforts will drive up the costs of open-ended performance assessments, creating pressures to return to multiple-choice tests. In any case, large-scale assessments cannot be expected to provide the quick turnaround that teachers require in order for test-based information to be useful in instructional decisions.

Although attempts to combine several functions in a single testing program are not advisable, performance assessments can also be used

for other kinds of educational functions, such as instructional diagnosis and selection. Selection testing, although it requires examination of individual students, needs to be done only once or twice in a student's educational career, keeping performance assessment costs within bounds. For college selection, the additional costs might even be included in the standard testing fee. In the case of diagnostic testing, it is perfectly acceptable — even desirable, in many cases — for students' own instructors to grade and interpret students' performances. Although this may take more of instructors' time than scoring multiple-choice tests with a machine, the time spent is directly relevant to the instructional process and should help to focus instructional efforts on the quality of students' thinking and reasoning.

Performance assessments are a feasible and attractive solution to the problems laid out in this essay. Properly developed and implemented, they allow for reliable measurement of thinking and reasoning in school subject matters. They offer a way to release educators from the pressure toward fractionated, low-level forms of learning rewarded by most current tests, and they can also set positive standards for an educational system that aims to cultivate thinking. Tied to curriculum and designed to be taught to, performance assessments can be essential tools for raising authentic educational achievement.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Blumberg, F., Epstein, M., MacDonald, W., & Mullis, I. (1986). *A pilot study of higher-order thinking skills assessment techniques in science and mathematics*. Princeton, NJ: National Assessment of Educational Progress.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16(8)*, 16-20.

Fuhrman, S. (1988). Educational indicators: An overview. *Phi Delta Kappan, 69(7)*, 486-487.

Gardner, H. (in press). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction*. Boston, MA: Kluwer Academic Publishers.

Joint Matriculation Board, Examinations Council. (1982). *General certificate of education: Engineering science (advanced) instructions and guidance for centres*. Manchester, England: Author.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1980). *The effects of standardized testing*. Dublin, Ireland/Boston, MA: St. Patrick's College/Boston College.

Lakatos, I. (1978). *The methodology of scientific research programmes, philosophical papers, Volume I*. New York, NY: Cambridge University Press.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18(2)*, 85-96.

Murnane, R. J., & Raizen, S. A. (Eds.). (1988). *Improving indicators of the quality of science and mathematics education in grades K-12*. Washington, DC: National Academy Press.

National Assessment of Educational Progress. (1987). *Learning by doing: A manual for teaching and assessing higher order thinking in science and mathematics*. Report 17-HOS-80. Princeton, NJ: Educational Testing Service.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Research Council. (1989). *Everybody counts -- A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1987). The merits of measurement driven instruction. *Phi Delta Kappan, 68(9),* 679-682.

Resnick, L. B. (1987a). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L. B. (1987b). Learning in school and out. *Educational Researcher, 16(9),* 13-20.

Resnick, L. B., Bill, V., & Lesgold, S. (September 1989). Developing thinking abilities in arithmetic class. Paper presented at the Third European Conference for Research on Learning and Instruction, Madrid.

Resnick, L. B., & Klopfer, L. E. (1989). Toward the thinking curriculum: An overview. In L. B. Resnick & L. E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research. 1989 ASCD Yearbook.* Alexandria, VA/ Hillsdale, NJ: Association for Supervision and Curriculum Development/ Erlbaum.

Romberg, T. A., Zarinnia, E. A., & Williams, S. (1989). The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions. Unpublished manuscript. Madison, WI: University of Wisconsin-Madison, National Center for Research in Mathematical Science Education.

Scribner, S. (1984). Studying working intelligence. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context.* Cambridge, MA: Harvard University Press.

Shepard, L. A. (April 1988). Should instruction be measurement driven: A debate. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Thorndike, E. L. (1922). *The psychology of arithmetic.* New York, NY: Macmillan.

Toulmin, S. E. (1972). *Human understanding.* Princeton, NJ: Princeton University Press.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power.* New York, NY: Basic Books.